Multi-view Neural Network Ensemble for Short and Mid-term Load Forecasting

Chun Sing Lai, *Senior Member, IEEE*, Yuxiang Yang, Keda Pan, Jianjun Zhang, *Student Member, IEEE*, Haoliang Yuan, *Member, IEEE*, Wing W. Y. Ng, *Senior Member, IEEE*, Ying Gao, Zhuoli Zhao, *Member, IEEE*, Ting Wang, *Student Member, IEEE*, Mohammad Shahidehpour, *Fellow, IEEE*, Loi Lei Lai, *Fellow, IEEE*

***Abstract*—Accurate load forecasting is essential to the operation and planning of power systems and electricity markets. In this paper, an ensemble of radial basis function neural networks (RBFNNs) is proposed which is trained by minimizing the localized generalization error for short-term and mid-term load forecasting. Exogenous features and features extracted from load series (with long short-term memory networks and multi-resolution wavelet transform) in various timescales are used to train the ensemble of RBFNNs. Multiple RBFNNs are fused as an ensemble model with high generalization capability using a proposed weighted fusion method based on the localized generalization error model. Experimental results on three practical datasets show that compared with other forecasting methods, the proposed method reduces the mean absolute percentage error (MAPE), mean squared error (MSE), mean absolute error (MAE) by at least 0.12%, 8.46 (MW)^2, 0.83 MW in mid-term load forecasting (i.e., to predict the daily peak load of next month), respectively, and reduces the MAPE, MSE by at least 0.19%, 2009.69 (MW)^2 and 0.30%, 3697.18 (MW)^2 in half-hour-ahead forecasting and day-ahead forecasting, respectively.**

***Index Terms*—Multi-view, ensemble, long short-term memory network, load forecasting**

Manuscript received March 19, 2020; revised July 30, 2020 and October 18, 2020; accepted November 29, 2020. Date of publication XXX; date of current version XXX. This work was supported by National Natural Science Foundation of China under Grants 61876066, 51907031, 61903091 and 61572201; Guangzhou Science and Technology Plan Project 201804010245; the Department of Finance and Education of Guangdong Province 2016 [202]: Key Discipline Construction Program, China; the Education Department of Guangdong Province: New and Integrated Energy System Theory and Technology Research Group [Project Number 2016KCXTD022]; Brunel Research Initiative and Enterprise Fund, UK. Paper no. TPWRS-XXXXX-XXXX. *(Corresponding authors: Y. Yang; M. Shahidehpour; L. L. Lai)*

C. S. Lai is with the Department of Electrical Engineering, Guangdong University of Technology, Guangzhou 510006, China, and also with the Brunel Institute of Power Systems, Department of Electronic and Electrical Engineering, Brunel University London, London UB8 3PH, U.K. (e-mail: chunsing.lai@brunel.ac.uk).

Y. Yang, J. Zhang, W. W. Y. Ng, Y. Gao, and T. Wang are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China (e-mail: fotonyoung@gmail.com; 754060224@qq.com wingng@ieee.org; gaoying@scut.edu.cn; tingwang@ieee.org)

M. Shahidehpour is with the Illinois Institute of Technology, Chicago, IL 60616 USA, and also with the Center of Research Excellence in Renewable Energy and Power Systems, King Abdulaziz University, Jeddah 21589, Saudi Arabia (e-mail: ms@iit.edu).

K. Pan, H. Yuan, Z. Zhao, and L. L. Lai are with the Department of Electrical Engineering, Guangdong University of Technology, Guangzhou 510006, China (e-mail: 1111904017@mail2.gdut.edu.cn; haoliangyuan@gdut.edu.cn; zhuoli.zhao@gdut.edu.cn; l.l.lai@ieee.org)

NOMENCLATURE

x_b	The b^{th} training sample
$S_Q(x_b)$	Local input space includes all unseen samples located near the training sample
S_Q	Union of all neighborhoods of the training sample
Q	A given real number
Δx	Stochastic perturbation
Δy	Output perturbations
R^*_{SM}	Upper bound of the localized generalization error
A	Difference between the maximum and the minimum values of target outputs
B	Maximum value of training mean square error
M	The number of hidden neurons
N	Number of training samples
R_{emp}	Training mean square error
$E_{SQ}((\Delta y)^2)$	Stochastic sensitivity measure of output differences
$p(\Delta x)$	Probability density function of the input perturbations
n_p	Number of base predictors
W_i	The weight of the i^{th} base predictor
$R^*_{SM}(i)$	The R^*_{SM} of the i^{th} base predictor
$\sum R^*_{SM}$	The sum of R^*_{SM} of base predictors
Y	The fusion result
w	The weight vector
O	The vector of forecast results of each base predictor

I. INTRODUCTION

LOAD forecasting provides efficient, reliable, and economical solutions for power system planning and

operation. Many operating decisions are based on load forecasts, including the scheduling and dispatch of generating units, reliability analyzes, security assessment, and maintenance outage planning in power systems [1-2]. Long-term load forecasting aims to assist in power system infrastructure planning, while mid-term and short-term load forecasting solutions are applied to power system operations [3-4]. The common timeframes for mid-term and short-term forecasting are weeks to few months ahead [5-6] and half-hour to few hours ahead [4], respectively.

With the rise of deregulation in many countries the opening of market competitions in electric power industries around the world, load forecasting has become an important tool in an energy management system which plays a key role in substantiating energy transactions in competitive electricity markets [7]. However, it has become more difficult to forecast hourly load demands, with the proliferation of variable energy resources, which introduces variability and nonstationarity into real-time loads. In practice, load forecast errors have demonstrated significant implications on profit margins, market shares, and shareholder values [8-9]. Accordingly, power system operators have to use reliable load forecasting results while taking the power system uncertainty into account as a key element in their decision-making process.

Many statistical methods have been used for load forecasting, including exponential smoothing method [10], regression analysis method [11], Kalman filter method [12], autoregressive integrated moving average (ARIMA) method [13], time-series techniques [14], and so on. These traditional methods are highly attractive because of their possible physical interpretations, mature technologies, and simple algorithms. However, they cannot properly represent the nonlinear behavior of the load series. Hence, artificial intelligence techniques including artificial neural networks (ANN) [15], support vector regression (SVR) [16], extreme learning machine (ELM) [17], probabilistic load forecasting method [18-19], and radial basis function neural network (RBFNN) [20-21] have been proposed for electricity load forecasting.

Ahmad et al. [22] proposed an accurate and fast converging short-term load forecasting model to improve the forecast accuracy for industrial applications in smart grid, in which the convergence rate of the forecasting strategy is enhanced by devising modifications in the ANN heuristic algorithm and training process. A novel ANN type reducer is proposed in [23] with the firing strengths of rules and their interval centroids to directly and optimally generate the defuzzied output of interval type-2 fuzzy logic system model for day-ahead load forecasting. Cecati et al. [20] investigated the effectiveness of some of the newest designed training algorithms including SVR, ELM, decay RBF neural networks (DRNNs), improved second-order algorithm, and error correction algorithm to train typical radial basis function (RBF) networks for short-term electricity load forecasting which exhibits good forecasting accuracy. Raza et al. [24] proposed an ensemble forecast framework with a systematic combination of three different predictors, namely feedforward neural network, Elman neural network and RBFNN, which are trained using the global particle swarm optimization to improve forecasting accuracies. The forecasting accuracy is significantly improved by combining these three predictors in an ensemble framework using a trim aggregation technique to combine the output of individual predictors.

In recent years, deep neural networks and deep learning methods have seen phenomenal success and have become one of the most effective methods in various fields, including electricity load and price forecasting. As one of the most powerful deep learning methods, the long short-term memory (LSTM) recurrent neural network has become a mature technology for processing sequence data and time series forecasting. The LSTM neural network is effective at capturing long-term temporal dependency features of historical electricity load series [4]. Several case studies were conducted to forecast the electricity load by using the LSTM neural network [25] which resulted in a significant improvement. An LSTM based hybrid probabilistic forecasting method was proposed for short-term load, wind, PV and price forecasting in power markets [26]. Shi et al. [27] first discussed potentials for employing state-of-the-art deep learning techniques for household short-term load forecasts with high uncertainty and volatility. The approach proposed a novel pooling-based deep recurrent neural network to address the overfitting challenges brought by naive deep networks. The networks enable the learning of spatial information shared among interconnected customers and hence allowing more learning layers before overfitting occurs.

Another useful load forecast technique, proposed in recent years, is wavelet transform (WT) [28]. Electric load series contain several nonstationary features such as trends, changes in level and slope, and seasonality [29]. These features are often the most important and challenging parts of the load signal which must be considered when dealing with nonstationarity. That is the motivation in using WT analysis to extract the nonstationary features is presented in this paper. The load series is divided into one low-frequency and some high-frequency subseries by the multi-resolution analysis of WT in the wavelet domain, with each subseries retaining a useful compromise between time domain and frequency domain information. These subseries are typically more stable with fewer outliers than the original load series and higher forecasting accuracies [29]. The number of decomposition levels may have a great impact on the performance of wavelet transformation. Different numbers of decomposition levels have been used in the literature, for example, two levels [27] and three levels [30].

Some discussions about the number of decomposition levels in the wavelet transform were presented in [31] which concluded that three levels of decomposition are a promising choice for load forecasting because load series can be described in a more thorough and meaningful way. Three levels of decomposition are proposed in this work. The reasons are twofold [24, 32]. Firstly, important high-frequency features of the original series are emphasized. Secondly, noisy parts of the series are separated. Using this new representation of the original load signal, one can build a model for load forecast whose inputs are based on information from both the original load sequence and the wavelet domain subseries [17]. Another method [32] is proposed to forecast load's future trend by independently forecasting subseries in the wavelet domain in which the final forecast is obtained by returning to the original domain (inverse transform).

In summary, the forecasting accuracy of the above methods can be enhanced. To enhance the diversity and performance of

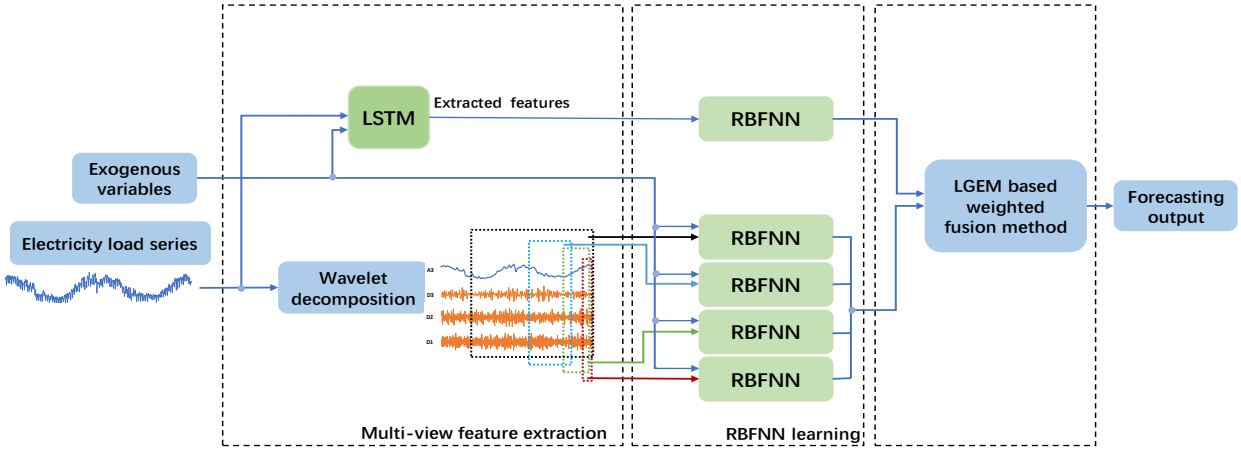


Fig. 1. Structure of the proposed ensemble forecasting framework.

each base predictor in an ensemble framework, and the accuracy of the forecasting model, a multi-view ensemble forecasting framework is proposed for electricity load forecasting. Major contributions of this work are summarized as follows:

1) A multi-view ensemble framework is proposed for electricity load forecasting with different views containing the knowledge extracted from the load series. Five views are proposed in the framework where one view is learned by using the LSTM network while the other four views are learned by using the wavelet decomposition. The view extracted by LSTM mainly contributes to the better learning of long-term recurrent patterns. The other four views are extracted by decomposing the original load series into four subseries, including an approximation component and three details components with different frequencies, each representing a specific view containing the time and frequency domain information. By combining these five views, the framework yields a high accuracy in load forecasting problems (i.e., with high uncertainty due to long-term recurrent events and/or short-term random variations).
2) For each wavelet decomposition subseries, the training dataset is divided into several different time window size datasets to obtain multivariate time series training sets. These different datasets represent the knowledge on different time periods about the historical load series and diversify the knowledge of predictors.
3) A weighted fusion method based on the localized generalization error model is proposed. Using this method, multiple RBFNNs trained by minimizing the localized generalization error are fused as an ensemble model with a high generalization capability for future unseen samples in electricity load forecasting.
4) The performance of the proposed load forecasting model is evaluated on three practical datasets. The proposed method yields the best results in terms of mean average percentage error (MAPE), mean square error (MSE), mean absolute error (MAE) and reduces the error from 0.12% to 0.98%, 8.46 $(MW)^2$ to 360.89 $(MW)^2$, 0.83 MW to 8.79 MW in mid-term load forecasting (i.e., to predict the daily peak load of next month), respectively, and reduces the MAPE, MSE from 0.19% to 1.56%, 2009.69 $(MW)^2$ to 21408.62 $(MW)^2$ and 0.30% to 2.85%, 3697.18 $(MW)^2$ to 166382.55 $(MW)^2$ in half-hour-ahead forecasting and day-ahead forecasting, respectively.

The rest of the paper is structured as follows. Section II introduces the ensemble framework and describes the whole procedure of the proposed method. Section III presents two case studies to evaluate the performance of the proposed method. The conclusion is given in Section IV.

II. Methodologies

In this paper, a multi-view ensemble load forecasting framework is proposed. The ensemble is constructed by fusing multiple radial basis function neural networks (RBFNN) which are trained with multi-view features extracted from both LSTM network and three-level wavelet decomposition. The RBFNNs are trained by minimizing the localized generalization error and fusion weights are determined according to their localized generalization error bounds. By employing the localized generalization error model (LGEM), the ensemble is expected to achieve higher generalization capability for future samples in electricity load series.

The structure of the proposed ensemble framework for load forecast is shown in Fig. 1. The ensemble framework consists of three main components: multi-view feature extraction, RBFNN-base predictors learning based on LGEM and the LGEM-based weighted fusion method. The details of the three components are presented in the following subsections.

A. Multi-view feature extraction

As aforementioned, five views containing different pieces of knowledge are extracted from the load series for enhancing the forecasting accuracy. The first view is extracted using the LSTM network while the other four views are extracted using three-level wavelet decomposition. The LSTM neural network is a kind of recurrent neural networks of which memory cells consist of an input gate that records the new information selectively into the cell state, an output gate that forgets the information selectively in cell states, and a forget gate that provides the results selectively as output, which is suitable for time series forecasting. As illustrated in Fig. 1, the LSTM network is used for feature extraction from the historical load series and exogenous variables.

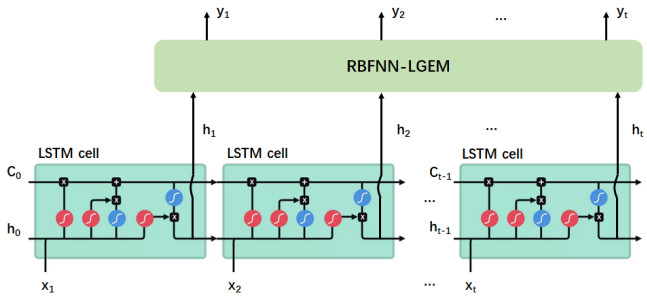


Fig. 2. The architecture of LSTM-RBFNN.

The LSTM neural network is effective at capturing long-term temporal dependency features of historical electricity load series [4]. As such, the view learned by LSTM mainly enhances long-term recurrent pattern learnings. The architecture of the proposed LSTM-RBFNN model is illustrated in Fig. 2. For the convenience of the readers, some details of LSTM's architecture are introduced in Appendix A. The LSTM network is used for feature extraction from the historical load series while the RBFNN is used for forecasting using the extracted feature of this view. The hidden layer of LSTM is a gated cell. It consists of four layers that interact with one another in a way to produce the output of that cell along with the cell state. These two pieces of information are then passed to the next hidden layer. LSTMs comprises of three logistic sigmoid gates and one tanh layer. Gates have been introduced in order to limit the information being passed through the cell. Given the gated architecture of LSTM that has this ability to manipulate its memory state, which is ideal for capturing the temporal feature in historical electricity load series. Besides, LSTM can almost seamlessly model problems with multiple input variables, where classical linear methods can be difficult to adapt to multivariate or multiple input forecasting problems. These features, together with several exogenous variables, are directly fed to train RBFNNs, which will be described in Section II.B below.

The extractions of the other four views, however, are different and are extracted by decomposing the original load series into four subseries, including an approximation component and three detail components with different frequencies by a three-level wavelet decomposition which uses the Symlet with 3 vanishing moments. The details of wavelet decomposition is introduced in Appendix B. A larger number of vanishing moments yields a more compact wavelet. After comparison, the number of vanishing moments is selected as 3. The decomposition retains a useful compromise between time and frequency domain information. The trend part (low frequency) is a smoothed version of the original load signal which captures the signal trend on a large timescale. First and second levels of variation parts (high frequencies) contain useful higher frequency information, which exhibits some regularities and similar shapes. Irregularities in these levels of detail are due to random load variations and measurement errors. The third level of variations shows some peaks that allow time localizations of peak load. These four subseries contain different views of information of the original load series, and using this information may help to improve the performance of the ensemble predictors.

After obtaining the four subseries, each subseries is combined with exogenous variables (which may differ in datasets) to construct a multivariate time series dataset. Training data for a long period of interest may originate inaccuracies because load characteristics vary with time. However, a short training period may originate volatile estimations. Therefore, a multiple time scale model is proposed to split the historical load series training dataset into several time window dataset, the feature in various timescales represent the knowledge on different periods of time about the problem and diversify the knowledge of predictors. Then, each of the four multivariate time series datasets is divided into multiple datasets by sliding window method. Each multivariate dataset employs a different window size. In this way, segmented datasets from different multivariate datasets present the information from different timescales. Each of these multiple timescales can now be used to train an RBFNN for load series forecast.

In summary, the first view is extracted using an LSTM network mainly for handling long-term recurrent events, while the other four views which are extracted using wavelet decomposition aim to handle patterns in both stationary (via trend part) and nonstationary environments (via variation parts). With these five views extracted from load series, trained predictors are expected to perform well in practical applications with large uncertainties.

B. RBFNN learning

RBFNN trained by minimizing the localized generalization error is used as the base predictor in the ensemble framework. Features extracted from five different views are fed to train RBFNNs, as illustrated in Fig. 1. These different features represent the knowledge on the time domain, frequency domain, seasonality characteristics information, and long-term temporal dependencies between local temporal patterns, and diversify the knowledge of predictors.

RBFNN is one of the most widely applied neural networks for pattern classification and prediction, with its performance primarily determined by its selected architecture. The generalization error is the most important criterion in the evaluation of a predictor. But the generalization error cannot be directly estimated because the whole input space can never be known accurately. Besides, it may be counter-productive to assess the generalization performance of the predictor on unseen samples that are very different from the training set. Hence it will be more sensible to develop a generalization error model for unseen samples which will be located within a neighborhood of training samples.

Yeung et al. [33] proposed an LGEM using the stochastic sensitivity measure (ST-SM), which bounds from above the generalization error for unseen samples within a predefined neighborhood of training samples. The purpose of RBFNN training by LGEM is to find a network structure and connection weight to minimize the localized generalization error. After the number of hidden neurons is determined, parameters such as center, width, and weight will be tuned by training samples. Therefore, the objective of RBFNN training can be simplified

to the optimal number of hidden neurons which makes use of the generalization capability of RBFNN.

The S_Q neighborhood is defined as the union of all $S_Q(x_b)$ neighborhoods of the training sample x_b, where x_b is the b^{th} training sample. The $S_Q(x_b)$ is defined as a local input space which includes all unseen samples located near the training sample x_b:

$$S_Q(x_b) = \{x | x = x_b + \Delta x, |\Delta x_i| \le Q, \ \forall i = 1,2,\ldots,n\} \quad (1)$$

where n is the number of input features, Q is a given real number, $x_b \in D$ where D is the training set, $\Delta x = (\Delta x_1, ..., \Delta x_n)$ denotes the stochastic perturbation and Δx_n denotes the stochastic perturbation on the n^{th} input feature.

For a given Q value, with a probability of $(1-\eta)$, the upper bound of the localized generalization error R^*_{SM} is estimated by using the Hoeffiding's inequality as follows:

$$\begin{aligned} R_{SM}(Q) &\le \frac{1}{N}\sum_{b=1}^{N}\int_{S_Q(X_b)} \big(f_\theta(x) - F(x)\big)^2 \frac{1}{(2Q)^n}\,dx + \varepsilon \\ &\le \left(\sqrt{R_{emp}} + \sqrt{E_{SQ}((\Delta y)^2) + A}\right)^2 + \varepsilon \\ &= R^*_{SM}(\mathrm{Q}) \end{aligned} \quad (2)$$

where $\varepsilon = B\sqrt{\ln\eta/(-2N)}$, N, A, B, R_{emp} and $E_{SQ}((\Delta y)^2)$ denote the number of training samples, difference between the maximum and the minimum values of target outputs, maximum value of training mean square error (MSE), training mean square error, and stochastic sensitivity measure (ST-SM) of output differences, respectively.

The output perturbation (Δy) measures the network output difference between the training sample ($x_b \in D$) and the unseen sample in its neighborhood ($(x_b + \Delta x) \in S_Q(x_b)$). Thus, ST-SM measures the expectation of the squares of network output perturbations (Δy) between training samples in D and unseen samples in S_Q. The ST-SM for an RBFNN is given by:

$$\begin{aligned} E_{SQ}((\Delta y)^2) = \frac{1}{N}\sum_{b=1}^{N}\int_{S_Q(X_b)} &\big(f_\theta(x_b + \Delta x) \\ &- f_\theta(x_b)\big)^2 p(\Delta x)\,d\Delta x \end{aligned} \quad (3)$$

where $p(\Delta x)$ denotes the probability density function of the input perturbations and $p(\Delta x) = 1/(2Q)^n$. By fixing Q, the optimal RBFNN is found by searching for the optimal number of hidden neurons which yields the minimum generalization error bound.

It is worth mentioning that the LGEM model and the neighborhood derived from Equations (2) and (3) are used to define the input space of the RBFNN instead of the original time series. The input spaces of RBFNNs are features extracted from different views which are of continuous nature. Particularly, the neighborhood in this work can be viewed as a set of unseen samples outside of training samples which has minor differences from the training samples. The differences or perturbations can be created by measurement error or round up in numbers (e.g. 36.1 from 36.132222). The differences are summarized as perturbations to the inputs. When extracted feature values are perturbed, it can be viewed as a perturbation, a measurement error or a new record with minor differences from the original time series. Therefore, even though it is a discrete time forecasting problem, the method has no issue to handle the solution.

The architecture selection algorithm of RBFNN is stated as follows:

1) Start with $M = 1$ (M denotes the number of hidden neurons).
2) Perform K-means clustering algorithm to find the centers for the M hidden neurons. K-means clustering partitions n observations into K clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster [34].
3) For each of the M RBF hidden neurons, select its width value as the distance between its center and that of the nearest hidden neurons.
4) Compute the connection weights using a pseudo inverse method.
5) Compute the R^*_{SM} for the current RBFNN.
6) If the stopping criterion is not fulfilled, set $M = M + 1$ and go to Step 2.

The stopping criterion is usually set as the number of hidden neurons to be equal to the number of training samples. However, in practical applications, the localized generalization error bound usually will not change when M is very large. As such, $M = 50$ is used in the experiments.

C. LGEM-based weighted fusion method

After training RBFNN, forecasting outputs of RBFNN-based predictors are aggregated by a weighted fusion method, which is based on localized generalization error bound to produce multi-view ensemble load forecasting results.

The ensemble fusion methods are mostly based on measures of diversity or performance of each base predictor. It is hard but necessary to keep a balance between diversity and performance. The base predictors with only high training accuracy may tend to produce similar results and thus may not improve the ensemble accuracy on unseen samples. In this work, LGEM is used to evaluate the performance of each base predictor. R^*_{SM} consists of three major components: training mean square error, stochastic sensitivity measure of output differences and the constants. If one classifier yields a smaller R^*_{SM} than the other, it means that the classifier will have a better generalization performance, so the base predictor with a lower R^*_{SM} value is supposed to be assigned with a higher weight because a base predictor yielding a smaller R^*_{SM} value has a lower generalization error. In addition, each base predictor is trained using features learned from different views for increasing the diversity. In this way, a multi-view ensemble with high generalization capability is obtained.

The weighted fusion method is used to aggregate base predictor outputs and the weight of a base predictor depends on its localized generalization error bound R^*_{SM}. The weight of the i^{th} base predictor is given as:

$$W_i = \begin{cases} \dfrac{\sum R^*_{SM} - (R^*_{SM}(i))}{\sum R^*_{SM} * (n_p - 1)}, & n_p > 1 \\ 1 & , n_p = 1 \end{cases} \quad (4)$$

where $\sum R^*_{SM}$ is the total R^*_{SM} of base predictors, $R^*_{SM}(i)$ is the R^*_{SM} of the i^{th} base predictor, and n_p is the number of base predictors. As shown in Equation (4), a smaller R^*_{SM} of a base classifier leads to a larger fusion weight. The final output of the ensemble network is given by:

$$\mathrm{Y} = w^T O \tag{5}$$

where Y denotes the fusion result, w is the weight vector, and O is the vector of forecast results of each base predictor.

III. Case Study

In this section, three case studies are conducted to evaluate the performance of the proposed method.

Section III.A reports on the month-ahead load forecasting results with a dataset from the EUNITE competition in 2001. The competition aims at predicting the daily maximum load on 31 days of Jan. 1999 using historical load data, daily temperature, and local holidays in 1997 and 1998 [35]. The exogenous variables in the EUNITE dataset include average daily temperature, temperature difference with the previous day, day of the week, and holiday information. There is a significant correlation between the maximum load and the average temperature in the EUNITE dataset, in which the load increases as the temperature drops in winter [36]. The existing methods for comparison are support vector machine (SVM) [36], local linear model trees (LoLiMOT) [37], SVM-Wavelet [38], Extended Bayesian [39], K-nearest neighbors (KNN) [40], ANN [41], ELM ensemble [42], and LSTM [4]. The corresponding results are provided in Table I and discussed in Section III.A. It should be noted that the results of SVM [36], LoLiMOT [37], SVM-Wavelet [38], and Extended Bayesian [39] are directly copied from the respective publications because they use the training and testing data from the competition as described in this work. Other results are produced by applying the corresponding method to the same training data and then the testing data.

Section III.B reports the results on half-hour-ahead and day-ahead forecasting using a dataset of the state Victoria from the Australian National Electricity Market (ANEM) [43]. The half-hour-ahead or day-ahead forecasting is to predict the next period load value with previously observed load records and relevant information. The historical half-hour load data of the state Victoria from the Australian Energy Market Operator (AEMO) are used for 2015 to 2018. In this case, exogenous variables consist of the month of the year, hour of the day, day of week and holiday information. In the experiments, the data for 2015 to 2017 is used for training and the data for 2018 is used for testing. The comparison methods for short-term load forecast are ELM ensemble [42], KNN [40], ANN [41], SVR, ARIMA and LSTM [4].

Section III.C reports the results on half-hour-ahead and day-ahead forecasting using a dataset of the state New South Wales from ANEM [43]. Compared to Section III.B, the exogenous variables in this dataset consist of the temperature information, hour of the day, day of week and holiday information. The temperature information including dry bulb, wet bulb, dew point temperatures and humidity, which are important to determine the state of humid air to have a significant correlation with electricity load demand [25]. Load demand increases in response to cooling needs in summer, and heating needs in winter. In the experiments, data of year 2006 to 2009 are used for testing. For this case study, the same comparison methods used in Section III.B are adopted.

For all datasets, binary data is used for representing the day of the week: 0 represents Monday to Friday and 1 represents Saturday and Sunday. Another binary number is used to represent holidays (i.e., special days) in which 0 represents non-holidays and 1 represents holidays.

The performance metrics employed in this work include MAPE and MSE, which are defined as:

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{A_i - P_i}{A_i}\right| \times 100\% \tag{6}$$

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(A_i - P_i)^2 \tag{7}$$

where A_i, $\bar{A}_i$ and P_i are actual, average actual, and forecasted load value on the i^{th} forecasted point respectively, and N is the number of forecasted points.

A. Results of month-ahead load forecasting based on EUNITE dataset

In this subsection, firstly a set of experiments is conducted to confirm the effectiveness of the combinations of different views which are learned from load series, i.e., a view extracted using the LSTM network and four views extracted using the wavelet decomposition. The exogenous variables (include average daily temperature, temperature difference with the previous day, day of the week, and holiday information) and historical load data are the input to the LSTM network, the exogenous variables (include average daily temperature, temperature difference with the previous day, day of the week, and holiday information) and historical load data are used as input features. The five input vectors are scaled to the range from 0 to 1 because the LSTM network is sensitive to data scale. Each row of the input matrix is the scaled features for the corresponding time step, which is fed to the corresponding LSTM block in the LSTM layer. In the month-ahead load forecasting, the window sizes are set to be a week, a month, a season and a year. After training the ensemble network, the forecasted temperature and other variables of the next month provided by the competition organizer are used to predict the daily maximum load demand of a month using the proposed model in a multi-step forecast. Then, the same dataset is used to confirm the efficacy of the proposed method by comparing it with existing methods. And these predictors are combined with some ensemble models (labeled as Ensemble of [40-41], Ensemble of [4, 40], Ensemble of [4, 41] and Ensemble of [4, 40-41]) for comparison. The base predictors in these ensemble methods are fused via the weighted fusion method using their MAPEs as fusion weights. The results of the proposed method compared with existing methods and their ensemble models are given in Table I.

It can be seen from Table I that the proposed method yields the best result with a MAPE of 1.60%. The second-best result is produced by the Ensemble of [4, 40] with a MAPE of 1.72%, and the best existing method is Extended Bayesian [39] which yields a MAPE of 1.75%. This shows that the proposed method improves the load forecasting performance by at least 0.12%. Besides, the proposed method yields the lower results in terms of MSE, MAE, standard deviation (SD) and reduces them by at least 8.46 $(\mathrm{MW})^2$, 0.83 MW and 0.35 MW, respectively.

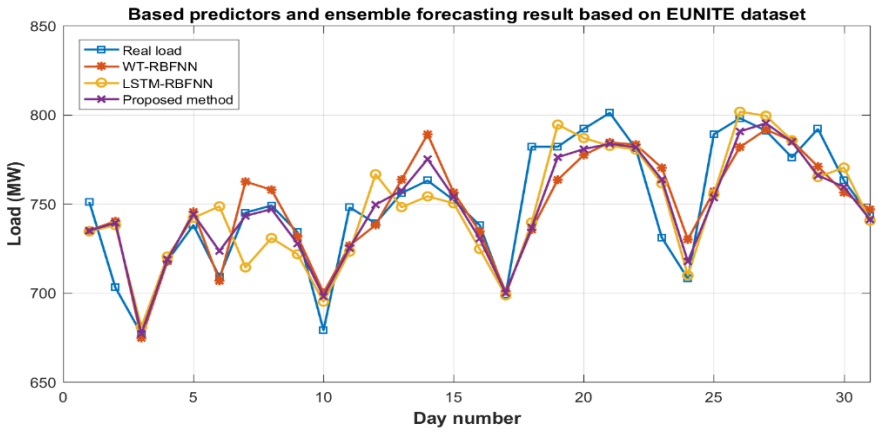


Fig. 3. Month-ahead real load and forecasted results based on EUNITE dataset

TABLE I
COMPARISON OF FORECASTING METHODS BASED ON EUNITE DATASET (MONTH-AHEAD FORECASTING)

Forecasting method	MAPE (%)	MSE ((MW)^2)	MAE (MW)	Standard deviation (SD) (MW)
SVM [36]	1.95	---	---	---
LoLiMOT [37]	1.98	---	---	---
SVM-Wavelet [38]	1.96	---	---	---
Extended Bayesian [39]	1.75	---	---	---
KNN [40]	2.48	567.87	18.34	31.70
ANN [41]	2.53	643.13	18.42	32.71
ELM ensemble [42]	2.58	628.50	18.73	33.71
LSTM [4]	2.05	406.83	15.71	31.85
Ensemble of [40-41]	2.38	539.63	17.56	31.45
Ensemble of [4, 40]	1.72	290.70	12.77	29.30
Ensemble of [4, 41]	1.78	309.06	13.27	30.27
Ensemble of [4, 40-41]	1.85	324.36	13.65	29.75
Proposed method	**1.60**	**282.24**	**11.94**	**28.95**

TABLE II
COMPARISON OF PROPOSED METHODS BASED ON EUNITE DATASET (MONTH-AHEAD FORECASTING)

Forecasting method	MAPE (%)	MSE ((MW)^2)	MAE (MW)	SD (MW)
WT-RBFNN	1.85	342.25	13.89	29.37
LSTM-RBFNN	2.01	381.81	15.06	31.60
Ensemble1	1.63	291.73	12.22	28.97
Ensemble2	1.62	289.68	12.17	28.88
Proposed method	**1.60**	**282.24**	**11.94**	**28.95**

The proposed ensemble framework consists of two sets of base predictors, one is trained using the features extracted from the LSTM (labeled as LSTM-RBFNN), and the other one uses the features from the views extracted by the wavelet decomposition (these predictors, labeled as WT-RBFNN, are also fused via the LGEM-weighted fusion method). The forecasting results of the proposed method compared with two components are shown in Fig. 3, the x-axis represents the day of Jan. 1999, and the y-axis represents the load values (in MW). It shows that the proposed method can better fit the curve of the real load series in almost all cases. To demonstrate the performance more clearly, the MAPE, MSE, MAE and SD of each method are computed and listed in Table II. It is seen that the accuracies of the proposed method and the other two compared ensemble model are acceptable, and the proposed ensemble model has the lowest MAPE , MSE and MAE. Ensemble2 yield a lower SD than the proposed method, which indicates that the forecasting result of the Ensemble2 are close to the mean set and dispersed at a narrower range, but the proposed method is supposed to be the most accurate forecasting method in this case study according to the MAPE, MSE, MAE performance metrics, which are the most commonly used key performance indicator to measure forecast accuracy. This proves the effectiveness of the combinations of different views and the LGEM based weighted fusion method helps to increase the forecast accuracy. In addition to this, the two components of the proposed method also achieved better performances against existing approaches in which WT-RBFNN performs better than LSTM-RBFNN. This implies that the views extracted by the three-level wavelet decomposition provide more useful information than LSTM in mid-term load forecasting.

To demonstrate the performance of different ensemble methods more clearly, the rest of Table II shows the comparative results of the proposed method with two ensemble frameworks. One compared ensemble framework (labeled as Ensemble1) consists of WT-RBFNN and LSTM-RBFNN base predictors which are fused via the average weighted fusion method. The other compared ensemble framework (labeled as Ensemble2) also consists of these two base predictors which are fused via the weighted fusion method using their MAPEs as fusion weights. The results of the proposed method compared with two compared ensemble frameworks are given in Table II. It is seen that the accuracies of the proposed method and the other two compared ensemble model are acceptable, and the proposed ensemble model can better fit the actual load, which shows that LGEM based weighted fusion method helps to increase the forecasting accuracy. It also should be noted that two compared ensemble frameworks (Ensemble1 and Ensemble2) also achieved better performances against two components and the Ensemble1 performs better than Ensemble2. This implies that the MAPE based weighted fusion method better aggregates the base predictors in mid-term load forecasting.

B. Half-hour-ahead and day-ahead load forecasting based on ANEM dataset of the state Victoria

This subsection focuses on short-term load forecasting. For this case study, a similar experimental setup is used as that in

TABLE III
COMPARISON OF SHORT-TERM FORECASTING METHODS BASED ON ANEM DATASET OF THE STATE VICTORIA

Forecasting method	Half-hour-ahead		Day-ahead	
	MAPE (%)	MSE ((MW)^2)	MAPE (%)	MSE ((MW)^2)
ELM ensemble [42]	2.53	27,369.87	7.27	268,987.45
KNN [40]	1.78	16,070.63	6.68	220,787.21
ANN [41]	2.86	31,314.84	6.17	165,893.29
LSTM [4]	1.49	11,915.91	4.95	129,945.83
SVR	2.42	25,998.34	5.78	181,493.04
ARIMA	1.72	20,198.09	5.42	142,589.31
Above methods ensemble	1.83	16,083.31	4.72	106,302.08
Proposed method	**1.30**	**9,906.22**	**4.42**	**102,604.90**

TABLE IV
COMPARISON OF PROPOSED SHORT-TERM FORECASTING METHODS BASED ON ANEM DATASET OF THE STATE VICTORIA

Forecasting method	Half-hour-ahead		Day-ahead	
	MAPE (%)	MSE((MW)^2)	MAPE (%)	MSE ((MW)^2)
WT-RBFNN	2.02	18,246.61	5.28	137,952.82
LSTM-RBFNN	1.42	11,569.15	4.55	103,317.24
Ensemble1	1.55	12,472.42	4.59	108,999.02
Ensemble2	1.35	10,342.89	4.51	105,690.01
Proposed method	**1.30**	**9,906.22**	**4.42**	**102,604.90**

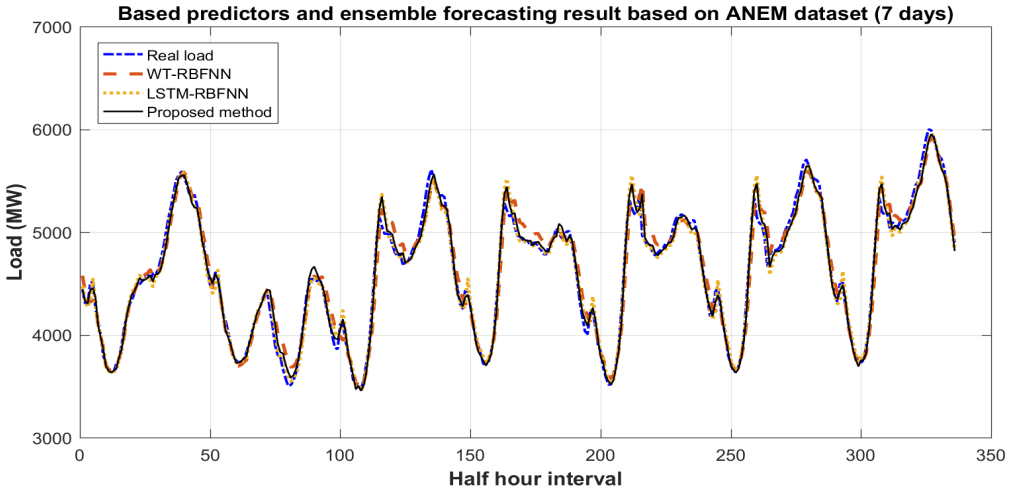


Fig. 4. Half-hour-ahead real load and forecasting result based on ANEM dataset

Section III.A. The results of the proposed method compared with existing methods in half-hour-ahead forecasting and day-ahead-forecasting are given in Table III. And these existing predictors are fused via the weighted fusion method using their MAPEs in the training set as fusion weights to ensemble models (labeled as Above methods ensemble) for comparison. In the half-hour-ahead forecasting, the proposed method yields the lowest MAPE of 1.30%, which is 0.19% better than that of the second-best method, i.e., LSTM [4] is with a MAPE of 1.49%. The proposed method yields the lowest MSE of 9,906.22 (MW)^2 as compared with the second-best MSE of 11,915.91 (MW)^2, i.e., LSTM [4]. Besides, in the day-ahead forecasting, the proposed method yields the lowest MAPE of 4.42%, which is 0.30% better than that of the second-best method, i.e., the Above methods ensemble, with a MAPE of 4.72%. The lowest MSE of 102,604.90 (MW)^2 is yielded by the proposed method as compared with the second-best MSE of 106,302.08 (MW)^2, i.e., the Above methods ensemble. Although the proposed method also yields the most accurate forecasting performance according to these performance metrics in day-ahead load forecasting, it is seen that the MAPE and MSE of both the proposed method and other comparison methods are high. The possible reason for this may be due to the lack of a detailed and accurate temperature data to make it more suitable for day-ahead forecasting.

The load forecasting results of the proposed method compared with the two components base predictors and two compared ensemble frameworks are given in Table IV. It can be concluded that the proposed method yields the most accurate forecasting performance according to these performance metrics. However, unlike the mid-term forecasting, WT-RBFNN performs worse than the LSTM-RBFNN in short-term forecasting. This outcome implies that the view extracted by LSTM contains more useful information than that extracted by the WT decomposition. The forecasting results of the proposed method and two components base predictors are illustrated in Fig. 4 which shows the actual load and the forecasting results in half-hour-ahead forecasting, where the x-axis represents the half-hour intervals in seven days in 2018, and the y-axis represents the load (in MW). It is seen that the accuracies of

TABLE V
COMPARISON OF SHORT-TERM FORECASTING METHODS BASED ON ANEM DATASET OF THE STATE NEW SOUTH WALES

Forecasting method	Half-hour-ahead		Day-ahead	
	MAPE (%)	MSE ((MW)^2)	MAPE (%)	MSE ((MW)^2)
ELM ensemble [42]	1.99	51,551.70	4.52	325,299.12
KNN [40]	1.73	44,276.58	4.26	300,095.80
ANN [41]	1.76	39,188.16	3.28	156,871.44
LSTM [4]	1.26	24,439.07	3.01	137,997.39
SVR	1.95	48,580.57	3.77	224,211.72
ARIMA	1.55	35,377.85	3.35	153,781.62
Above methods ensemble	1.25	24,152.27	2.81	130,834.12
Proposed method	**1.09**	**17,947.96**	**2.57**	**103,458.72**

TABLE VI
COMPARISON OF PROPOSED SHORT-TERM FORECASTING METHODS BASED ON ANEM DATASET OF THE STATE NEW SOUTH WALES

Forecasting method	Half-hour-ahead		Day-ahead	
	MAPE (%)	MSE((MW)^2)	MAPE (%)	MSE ((MW)^2)
WT-RBFNN	1.32	26,529.89	3.16	130,848.59
LSTM-RBFNN	1.21	22,873.54	2.89	129,297.78
Ensemble1	1.16	19,099.24	2.66	105,449.57
Ensemble2	1.15	18,950.28	2.61	103,767.74
Proposed method	**1.09**	**17,947.96**	**2.57**	**103,458.72**

both the proposed method and two components base predictors are acceptable, and the proposed ensemble model can better represent the actual load.

The results in Table IV also implies that a combination of LSTM and the three-level WT decomposition enables the proposed method and two ensemble frameworks to perform better than two components for both half-hour-ahead and day-ahead load forecasting. Same as the mid-term forecasting, Ensemble2 performs worse than the Ensemble1, and the proposed method yields superior performance in comparison to Ensemble1 and Ensemble2 for both mid-term and short-term load forecasting.

C. Half-hour-ahead and day-ahead load forecasting based on ANEM dataset of the state New South Wales

This subsection focuses on short-term load forecasting based on ANEM dataset of the state of New South Wales which contains a detailed temperature information data. For this case study, similar experimental setup and comparison methods as in Section III.A and Section III.B are used. The results of the proposed method compared with existing methods in half-hour-ahead forecasting and day-ahead-forecasting are given in Table V and the load forecasting results of the proposed method compared with two components base predictors and two compared ensemble frameworks are given in the Table VI. It can be concluded that the proposed method yields the most accurate forecasting performance according to these performance metrics.

As seen from Table V, in the half-hour-ahead forecasting, the proposed method yields the lowest MAPE of 1.09%, which is 0.16% better than that of the second-best method (Above methods ensemble) with a MAPE of 1.25% and the best existing method is LSTM [4] which yields a MAPE of 1.26%. The proposed method yields the lowest MSE of 17947.96 (MW)^2 as compared with the second-best MSE of 24152.27 (MW)^2, i.e., the Above methods ensemble. Similarly, in the day-ahead forecasting, the proposed method yields the lowest MAPE of 2.57%, which is 0.24% better than that of the second-best method (Above methods ensemble) with a MAPE of 2.81% and the best existing method is LSTM [3] which yields a MAPE of 3.01%. The proposed method yields the lowest MSE of 103,458.82 (MW)^2 as compared with the second-best MSE of 130,834.12 (MW)^2, i.e., the Above methods ensemble. Unlike the case study in Section III.B, the MAPE and MSE of both the proposed method and other comparison methods are much lower than the MAPE and MSE yielded in Section III.B in the day-ahead load forecasting. The possible reason for this superior performance may be due to this dataset includes the temperature data (dry bulb, wet bulb, dew point temperatures and humidity) in this case study which helps increasing the forecasting accuracy in the day-ahead load forecasting.

Table V reports the results of the proposed method against two components base predictors and two compared ensemble frameworks both in the day-ahead and half-hour-ahead forecasting. Similarly, the outcome in Table IV implies that a combination of LSTM and the three-level WT decomposition enables the proposed method and two ensemble frameworks to perform better than two components for both half-hour-ahead and day-ahead load forecasting, and the proposed method yields superior performance in comparison to Ensemble1 and Ensemble2 for both mid-term and short-term load forecasting.

In summary, the proposed ensemble load forecasting framework yields superior performance in comparison to existing methods for mid-term and short-term load forecasting. The possible reason for the effectiveness of the proposed method may be due to different views extracted from load series that are properly utilized and thus can enhance the performance of the predictors. Another reason is that LGEM helps to increase the generalization capability of the base predictor by selecting the optimal number of hidden neurons and ensemble by weighting base predictors using their localized generalization error bound.

IV. Conclusion

In this paper, a multi-view ensemble framework is proposed for enhancing the load forecasting results. Multi-views of features are first extracted from both the LSTM network and the three-level wavelet decomposition, which capture the characteristics of long-term recurrent events, trends, and variations of load series. These features, together with some exogenous variables, are then used to train the base predictors, RBFNN. To improve the generalization capability of the base predictor, LGEM is employed to search for the optimal network architecture. After learning the base predictors, these selectors are fused using a weighted fusion method where the weight of each predictor is determined by its own localized generalization error bound. In this way, the ensemble is expected to achieve a highly generalized performance for future data in load series. Experiments on three practical datasets confirm the effectiveness of the proposed ensemble framework. Future work will include applying the framework to other time series forecasting problems, including solar and wind power forecasting.

References

[1] L. Wu and M. Shahidehpour, "A Hybrid Model for Integrated Day-ahead Electricity Price and Load Forecasting in Smart Grid," *IET Generation Transmission and Distribution*, vol. 8, no. 12, pp. 1937-1950, 2014.

[2] O. Abedinia, N. Amjady and H. Zareipour, "A New Feature Selection Technique for Load and Price Forecast of Electrical Power Systems," *IEEE Transactions on Power Systems*, vol. 32, no. 1, pp. 62-74, 2017.

[3] B. F. Huang, D. Q. Wu, C. S. Lai, X. Cun, H. L. Yuan, F. Y. Xu, L. L. Lai and K. F. Tsang, "Load Forecasting based on Deep Long Short-term Memory with Consideration of Costing Correlated Factor," *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)*, Portugal, pp. 496 – 501, 2018.

[4] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu and Y. Zhang, "Short-term Residential Load Forecasting Based on LSTM Recurrent Neural Network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841-851, 2019.

[5] D. Ali, M. Yohanna, P. M. Ijasini and M. B. Garkida, "Application of Fuzzy–Neuro to Model Weather Parameter Variability Impacts on Electrical Load based on Long-term Forecasting," *Alexandria Engineering Journal*, vol. 57, no. 157, pp. 223-233, 2018.

[6] W. Jiang, H. Tang, L. Wu, H. Huang and H. Qi, "Parallel Processing of Probabilistic Models-Based Power Supply Unit Mid-Term Load Forecasting with Apache Spark," *IEEE Access*, vol. 7, pp. 7588-7598, 2019.

[7] S. Fan and R. J. Hyndman, "Short-term Load Forecasting based on a Semi-Parametric Additive Model," *IEEE Transactions on Power Systems*, vol. 27, no. 1, pp. 134-141, 2012.

[8] Y. Hsiao, "Household Electricity Demand Forecast based on Context Information and User Daily Schedule Analysis from Meter Data," *IEEE Transactions on Industrial Informatics*, vol. 11, no. 1, pp. 33-43, 2015.

[9] B. Stephen, R. Telford and S. Galloway, "Non-Gaussian Residual based Short Term Load Forecast Adjustment for Distribution Feeders," *IEEE Access*, vol. 8, pp. 10731-10741, 2020.

[10] A. Garulli, S. Paoletti and A. Vicino, "Models and Techniques for Electric Load Forecasting in the Presence of Demand Response," *IEEE Transactions on Control Systems Technology*, vol. 23, no. 3, pp. 1087-1097, 2015.

[11] J. Zhao and X. Liu, "A Hybrid Method of Dynamic Cooling and Heating Load Forecasting for Office Buildings based on Artificial Intelligence and Regression Analysis," *Energy and Buildings*, vol. 174, pp. 293-308, 2018.

[12] Z. Zheng, H. Chen and X. Luo, "A Kalman Filter-based Bottom-up Approach for Household Short-term Load Forecast," *Applied Energy*, vol. 250, pp. 882-894, 2019.

[13] M. S. Al-Musaylh, R. C. Deo, J. F. Adamowski and Y. Li, "Short-term Electricity Demand Forecasting with MARS, SVR and ARIMA Models using Aggregated Demand Data in Queensland, Australia," *Advanced Engineering Informatics*, vol. 35, pp. 1-16, 2018.

[14] H. J. Sadaei, P. C. D. L. e Silva, F. G. Guimarães and M. H. Lee, "Short-term Load Forecasting by using a Combined Method of Convolutional Neural Networks and Fuzzy Time Series," *Energy*, vol. 175, pp. 365-377, 2019.

[15] F. Y. Xu, X. Cun, M. Yan, H. Yuan, Y. Wang and L. L. Lai, "Power Market Load Forecasting on Neural Network with Beneficial Correlated Regularization," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 5050-5059, 2018.

[16] J. Che and J. Wang, "Short-term Load Forecasting using a Kernel-based Support Vector Regression Combination Model," *Applied Energy*, vol. 132, pp. 602-609, 2014.

[17] S. Li, P. Wang and L. Goel, "Short-term Load Forecasting by Wavelet Transform and Evolutionary Extreme Learning Machine," *Electric Power Systems Research*, vol. 122, pp. 96-103, 2014.

[18] M. Pierro, M. De Felice, E. Maggioni, D. Moser, A. Perotto, F. Spada and C. Cornaro. "Residual Load Probabilistic Forecast for Reserve Assessment: A Real Case Study," *Renewable Energy*, vol. 149, pp. 508-522, 2020.

[19] Y. Wang, N. Zhang, Y. Tan, T. Hong, D. S. Kirschen and C. Kang, "Combining Probabilistic Load Forecasts," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3664-3674, 2019.

[20] C. Cecati, J. Kolbusz, P. Różycki, P. Siano and B. M. Wilamowski, "A Novel RBF Training Algorithm for Short-Term Electric Load Forecasting and Comparative Studies," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 10, pp. 6519-6529, 2015.

[21] O. Abedinia and N. Amjady. "Short-term Load Forecast of Electrical Power System by Radial Basis Function Neural Network and New Stochastic Search Algorithm," *International Transactions on Electrical Energy Systems,* vol. 26, pp. 1511-1525, 2016.

[22] A. Ahmad, N. Javaid, M. Guizani, N. Alrajeh and Z. A. Khan, "An Accurate and Fast Converging Short-term Load Forecasting Model for Industrial Applications in a Smart Grid," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2587-2596, 2017.

[23] A. Khosravi and S. Nahavandi, "Load Forecasting Using Interval Type-2 Fuzzy Logic Systems: Optimal Type Reduction," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1055-1063, 2014.

[24] M. Q. Raza, M. Nadarajah, J. Li and K. Y. Lee, "Multivariate Ensemble Forecast Framework for Demand Prediction of Anomalous Days," *IEEE Transactions on Sustainable Energy*, vol. 11, no. 1, pp. 27-36, 2020.

[25] F. He, J. Zhou, Z. Feng, G. Liu and Y. Yang "A Hybrid Short-term Load Forecasting Model based on Variational Mode Decomposition and Long Short-term Memory Networks considering Relevant Factors with Bayesian Optimization Algorithm," *Applied Energy*, vol. 237, pp. 103-116, 2019.

[26] J. Toubeau, J. Bottieau, F. Vallée and Z. De Grève, "Deep Learning-based Multivariate Probabilistic Forecasting for Short-term Scheduling in Power Markets," *IEEE Transactions on Power Systems*, vol. 34, pp. 1203-1215, 2019.

[27] H. Shi, M. Xu and R. Li, "Deep Learning for Household Load Forecasting—A Novel Pooling Deep RNN," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 5271-5280, 2018.

[28] N. G. Paterakis, A. Taşcıkaraoğlu, O. Erdinç, A. G. Bakirtzis and J. P. S. Catalão, "Assessment of Demand-response-driven Load Pattern Elasticity using a Combined Approach for Smart Households," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 4, pp. 1529-1539, 2016.

[29] H. Chitsaz, N. Amjady and H. Zareipour, "Wind Power Forecast using Wavelet Neural Network Trained by Improved Clonal Selection Algorithm," *Energy conversion and Management*, vol. 89, pp. 588-598, 2015.

[30] X. Tong, C. Kang and Q. Xia, "Smart Metering Load Data Compression Based on Load Feature Identification," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2414-2422, 2016.

[31] N. Amjady and F. Keynia, "Day Ahead Price Forecasting of Electricity Markets by a Mixed Data Model and Hybrid Forecast Method," *International Journal of Electrical Power & Energy Systems*, vol. 30, pp. 533-546, 2008.

[32] M. Rafiei, T. Niknam, J. Aghaei, M. Shafie-Khah and J. P. S. Catalão, "Probabilistic Load Forecasting Using an Improved Wavelet Neural Network Trained by Generalized Extreme Learning Machine," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6961-6971, 2018.

[33] D. S. Yeung, W. W. Y. Ng, D. Wang, E. C. C. Tsang and X. Wang, "Localized Generalization Error Model and its Application to Architecture Selection for Radial Basis Function Neural Network," *IEEE Transactions on Neural Networks*, vol. 18, no. 5, pp. 1294-1305, 2007.

[34] C. S. Lai, Y. Jia, M. D. McCulloch and Z. Xu, "Daily Clearness Index Profiles Cluster Analysis for Photovoltaic System," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2322-2332, 2017.

[35] The EUNITE competition. [Online]. Available: https://rdrr.io/cran/TSPred/man/EUNITE.Loads.html.

[36] B. J. Chen, M. W. Chang, and C. J. Lin, "Load Forecasting using Support Vector Machines: A Study on EUNITE Competition 2001," *IEEE Transactions on Power Systems*, vol. 19, pp. 1821-1830, 2004

[37] A.R.Koushki, M. Nosrati Maralloo, C. Lucas and A. Kalhor, "Application of Neuro-Fuzzy Model in Short Term Electricity Load Forecast," *Proceeding of the 14th International CSI Computer Conference* 2009, pp. 41-46, 2009.

[38] J. Xu, X. Qiu and Z. Zhang, "The Application of Two New Integrated Models in Short-term Load Forecast," *2010 Asia-Pacific Power and Energy Engineering Conference*, Chengdu, pp. 1-4, 2010.

[39] V. H. Ferreira and A. P. A. da Silva, "Toward Estimation Autonomous Neural Network-based Electric Load Forecasters," *IEEE Transactions on Power Systems*, vol. 22, pp. 1554-1562, 2007.

[40] R. Zhang, Y. Xu, Z. Y. Dong, W. Kong and K. P. Wong, "A Composite k-nearest Neighbor Model for Day-ahead Load Forecasting with Limited Temperature Forecasts," *2016 IEEE Power and Energy Society General Meeting (PESGM)*, Boston, MA, pp. 1-5, 2016.

[41] V. Dehalwar, A. Kalam, M. L. Kolhe and A. Zayegh, "Electricity Load Forecasting for Urban Area using Weather Forecast Information," *2016 IEEE International Conference on Power and Renewable Energy (ICPRE)*, Shanghai, pp. 355-359, 2016.

[42] R. Zhang, Z. Y. Dong, Y. Xu, K. Meng and K. P. Wong. "Short-term Load Forecasting of Australian National Electricity Market by an Ensemble Model of Extreme Learning Machine," *IET Generation, Transmission and Distribution*. vol. 7(4), pp. 391-397, 2013.

[43] The ANEM dataset. [Online]. Available: https://www.aemo.com.au/Electricity/National-Electricity-Market-NEM/Data-dashboard#aggregated-data.

APPENDIX A: BACKGROUND OF LSTM NETWORK

The architecture of LSTM cell is presented in Fig. A. The LSTM cells consist of an input gate that records the new information selectively into the cell state, an output gate that forgets the information selectively in cell states, and a forget gate that provides the results selectively as output. First, the previous hidden state h_{t-1} and the current input x_t get concatenated, and the combined result is fed into the forget layer, candidate layer, and the input layer. The forget layer decides what information should be thrown away or kept, the combined information is passed through a sigmoid function. Values are scaled to the range from 0 to 1; the closer to 0 means to forget, and the closer to 1 means to keep. The candidate holds possible values to add to the cell state. The input layer decides what data from the candidate should be added to the new cell state by multiplying the tanh output with the sigmoid output. After computing the forget layer, candidate layer, and the input layer, the cell state c_t is calculated using those vectors and the previous cell state c_{t-1}. The output is then computed, and the new hidden state h_t is computed by multiplying the pointwise output and new cell state. Combining all those mechanisms, an LSTM network can choose which information is relevant to remember or forget during the sequence processing.

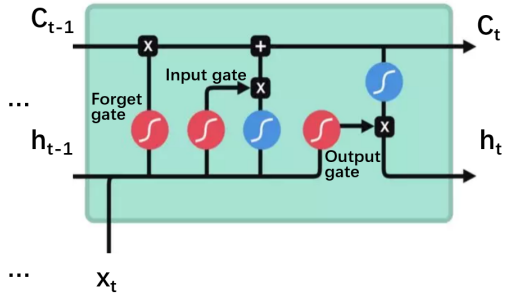


Fig. A. The architecture of a LSTM cell

APPENDIX B: WAVELET DECOMPOSITION

Wavelet Transform (WT) is carried out by the means of a special analyzing function ψ, called the basic wavelet. The continuous wavelet transform is defined as follows:

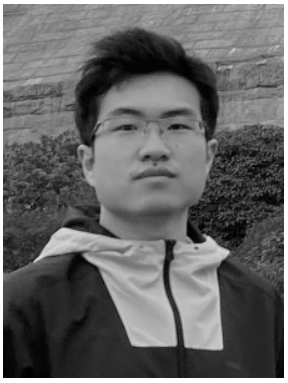
$$\mathrm{WT}(s,\tau) = \frac{1}{\sqrt{\tau}} \int_{-\infty}^{\infty} f(t) * \psi\left(\frac{t-\tau}{s}\right) dt \qquad \text{(B)}$$

As seen in the above equation, the transformed signal is a function of two variables, τ and s, the translation and scale parameters, respectively. During the analysis, this wavelet is translated in time (for selecting the part of the signal to be analyzed), then dilated/expanded or contracted/compressed using a scale parameter s (in order to focus on a given range or number of oscillations). When the wavelet is expanded, it focuses on the signal components which oscillate slowly (i.e., low frequencies); when the wavelet is compressed, it observes the fast oscillations (i.e., high frequencies), this is similar to those contained in a discontinuity of a signal. In brief, the WT is a correlation between a wavelet at different scales and the signal with the scale (or the frequency) being used as a measure of similarity. The continuous wavelet transform was computed by changing the scale of the analysis window, shifting the window in time, multiplying by the signal, and integrating over all times. In the discrete case, filters of different cut-off frequencies are used to analyze the signal at different scales. The signal is passed through a series of high pass filters to analyze the high frequencies, and it is passed through a series of low pass filters to analyze the low frequencies.

A time series data whose frequency content does not change in time is called stationary data. In other words, the frequency content of stationary signals does not change in different time. The frequency spectrum of the load series can be obtained by Fourier Transform (FT), which shows what frequencies exist in the signal, but it cannot tell when these frequency components exist. This is why FT is not suitable if the data has time varying frequency (i.e., the data is non-stationary). However, WT is capable of providing the time and frequency information simultaneously, hence giving a time-frequency representation of the load series which gives the information regarding when those frequency spectrum components appear in the non-stationary data.

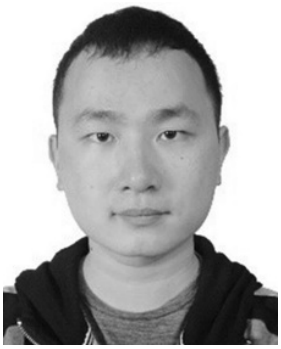
Chun Sing Lai (S'11, M'19, SM'20) received the B.Eng. (First Class Honours) in electrical and electronic engineering from Brunel University London, UK and D.Phil. in engineering science from the University of Oxford, UK in 2013 and 2019, respectively. Dr Lai is currently a Lecturer at Department of Electronic and Electrical Engineering, Brunel University London, UK and also a Visiting Academic with the Department of Electrical Engineering, Guangdong University of Technology, China. From 2018 to 2020, he was an Engineering and Physical Sciences Research Council Research Fellow with the School of Civil Engineering, University of Leeds.

He is Secretary of the IEEE Smart Cities Publications Committee and Acting EiC of IEEE Smart Cities Newsletters. He was the Publications Co-Chair for 2020 IEEE International Smart Cities Conference. Dr Lai is the Working Group Chair for IEEE P2814 Standard. His current research interests are in power system optimization and data analytics.



Yuxiang Yang received the B.Sc. and M.Sc. degrees from the South China University of Technology, Guangzhou, China, in 2015 and 2018, respectively. He is currently pursuing a Ph.D. degree with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His current research interests include neural networks, deep learning, and their applications in smart grid.

Keda Pan received the B.S. degree in electrical engineering from North China Electric Power University, Beijing, China, in 2015, and the M.S. degree in electrical engineering from Guangdong University of Technology, Guangdong, China, in 2019, respectively. He is currently working towards a Ph.D. degree at Guangdong University of Technology, Guangdong, China. His research interests include data analytics in smart grid and renewable energy integration.



Jianjun Zhang (S'19) received the bachelor's degree in computer science from the South China University of Technology, Guangzhou, China, in 2015, where he is currently working towards a Ph.D. degree in computer science with the School of Computer Science and Engineering. His current research interests include neural networks and machine learning for imbalanced and nonstationary environments.

Haoliang Yuan (Member) received the B.Eng. and M.Sc. degrees from the Hubei University, Wuhan, China, in 2009 and 2012, and the Ph.D. degree from the University of Macau, 2016. He is an Associate Professor in the School of Automation, Guangdong University of Technology.

Wing W. Y. Ng (S'02–M'05–SM'15) received the B.Sc. and Ph.D. degrees in computer science from Hong Kong Polytechnic University, Hong Kong, in 2001 and 2006, respectively.

He is a Professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He is currently the Deputy Director of the Guangdong Provincial Key Laboratory of the Computational Intelligence and Cyberspace Information Guangzhou, China. He is the Principle Investigator of four China National Nature Science Foundation Projects and a Program for New Century Excellent Talents in University from the Ministry of Education, China. His research interests include neural networks, deep learning, smart grid, smart health care, smart manufacturing, and nonstationary information retrieval.

Dr. Ng is an Associate Editor of the International Journal of Machine Learning and Cybernetics. He served as the Board of Governor of IEEE Systems, Man and Cybernetics Society in 2011–2013.

Ying Gao received the bachelor's degree in information engineering and the master's degree in computer application technology from Central South University, Changsha, China, in 1997 and 2000, respectively, and the Ph.D. degree in computer science and technology from South China University of Technology, Guangzhou, China, in 2006.

She is currently a Professor of Computer Science with the School of Computer Science and Engineering, South China University of Technology. She has published more than 30 papers in international journals and conferences. Her current research interests include service-oriented computing technology, software architecture and network security.

Zhuoli Zhao (S'15–M'18) received the Ph.D. degree in electrical engineering from South China University of Technology, Guangzhou, China, in 2017. From October 2014 to December 2015, he was a Joint Ph.D. Student (Sponsored Researcher) with the Control and Power Research Group, Department of Electrical and Electronic Engineering, Imperial College London, London, UK. He was a Research Associate with the Smart Grid Research Laboratory, Electric Power Research Institute, China Southern Power Grid, Guangzhou, China, from 2017 to 2018.

He is currently an Associate Professor with the School of Automation, Guangdong University of Technology, Guangzhou, China.

His research interests include microgrid control and energy management, power electronic converters, smart grids, and distributed generation systems. He is an Active Reviewer for the IEEE Transactions on Smart Grid, IEEE Transactions on Power Electronics, IEEE Transactions on Sustainable Energy, IEEE Transactions on Industrial Electronics, and Applied Energy.

Ting Wang (Student Member) received the M.Sc. degree in computer science from Northeast Normal University, Changchun, China, in 2017. She is currently pursuing a Ph.D. degree with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China.

Her current research interests include learning methods and generalization capabilities for deep neural networks, and their applications in real-world problems, such as smart healthcare and smart grid.

Mohammad Shahidehpour (F'01) received the Honorary Doctorate degree from the Polytechnic University of Bucharest, Bucharest, Romania. He is a University Distinguished Professor and serves as the Director of the Robert W. Galvin Center for electricity innovation, Illinois Institute of Technology. He is also a Research Professor with the Center of Research Excellence in Renewable Energy and Power Systems, King Abdulaziz University, Jeddah, Saudi Arabia. He was the recipient of the 2019 IEEE PES Ramakumar Family Renewable Energy Excellence Award. He is a member of the US National Academy of Engineering.

Loi Lei Lai (M'87, SM'92, F'07) received B.Sc. (First Class Honours) in 1980, Ph.D. in 1984 and D.Sc. in 2005 from University of Aston, UK and City, University of London, UK, respectively, all in Electrical and Electronic Engineering.

Prof Lai is University Distinguished Professor at the Guangdong University of Technology, China. He was Pao Yue Kong Chair Professor, Zhejiang University, China; Director of Research and Development Centre, State Grid Energy Research Institute, China; Vice President for IEEE SMC Society; Professor & Chair in Electrical Engineering at City, University of London; and a Fellow Committee Evaluator for IEEE IES. He was awarded IEEE PES UKRI Power Chapter Outstanding Engineer Award in 2000, IEEE PES Energy Development and Power Generation Committee Prize Paper in 2006 & 2009, IEEE SMCS Most Active Technical Committee Award in 2016; his research team has received a best paper award from the 2020 IEEE International Smart Cities Conference. He is a Fellow of IET. His current research areas are in smart cities and smart grid.